RESEARCH FOR SUPPORTING JOINT VENTURE USING ORGANIZATION INFORMATION ON WEB

Masanori Ikebe [1], Shigenori Tanaka [2], Hitoshi Furuta [3], Kenji Nakamura [4], and Kenta Kobayashi [5]

ABSTRACT

Today, the case has increased that some company organizes JV and takes part in the construction works because it's difficult for a single company to correspond to a large-scale construction works. However, there's a problem that a lot of JV have only been organized among the companies connected at the past. This is the reason that each company cannot share the reliable information such as its technology, capital and so on with each other. Then, this paper suggests how to analyze the organization information from the Web and extract the reliable relations among the companies. We suggest the Web structure analysis that combined link structure analysis with the natural language processing in order to collecting the reliable information. We adopt the improved HITS algorithm as the link structure analysis. And, in the natural language processing, we extract any topic information using the morphological analysis and analyze the information related to the Web pages without link relations. Then, in order to confirm the suggested method available, we compare the result of the method with the fact that is united JV using F measure that is used to judge the precision of the extracted information. We confirm that the method is available.

KEY WORDS

link structure analysis, morphological analysis, information collecting, graph theory

INTRODUCTION

The late years, the engineering works business becomes complicated and diversifies as the demand for the engineering works administration is increased and the technical development is progressed. Then, the high technologies for some fields are needed to run the smooth business and the effective administration because it is difficult for a single company to correspond to a large-scale execution (Ministry of Land, Infrastructure and Transport

[1] Ph.D Candidate, Master of Informatics, Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, ikebe@kansai-labo.co.jp

[2] Professor, Dr. Eng., Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, tanaka@res.kutc.kansai-u.ac.jp

[3] Professor, Dr. Eng., Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, furuta@res.kutc.kansai-u.ac.jp

[4] Master's Course Student, Bachelor of Informatics, Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, nakamura@kansai-labo.co.jp

[5] Under Graduate Course Student, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, kobayashi@kansai-labo.co.jp

2005). There are a lot of companies that considers cooperating with the businesses for diversifying their technology. There are various cooperation forms such as business merger, technology tie-up and joint enterprises. This cooperation shows several organization forms. And, it is necessary for suitable cooperation among the companies to be clear the several processes. In case of the cooperating among the companies, first, one company looks for a connection targeted company and makes a survey of its result, management situation, proud fields and related companies. Second, after deciding the targeted company, they confirm common recognitions and decide a cooperation policy. And they make sure the cooperation results and obstacles. Last, they have the concrete talks and realize the cooperation. Especially, it is important and difficult for the efficient cooperation to survey the target companies (Miyazaki Prefecture 2005). Then, as a method to help business analysis, an existing research devises a balance score card to analyze an achievement of companies (Jiro Shibano 2004) and a technique to evaluate whether business activity is smoothly done from the communication data (Fusashi Nakamura and Hideyuki Mizuta 2004). However, both researches need the detailed information about the targeted company. Moreover, they subjectively need to choose the targeted companies beforehand. In civil engineering works and construction industry community, Joint Venture (JV) appeared as a form of new connections. JV is a cooperation that some companies take part in the construction when it is difficult for a single company in the fund, technology, work force and risk of execution to correspond to a large-scale construction. By this connection, each company is able to make use of the special fields. And thanks to enforcing the construction jointly, the participated companies can acquire new knowledge (Chen Chung-Jen 2004), (Alice Nakamura and Masao Nakamura 2004). However, 53% JV (James Bamford et al 2004) has problems such as a want of reliability and an uncertain strategy due to the lack of investigation and communication. Then, the existing research (Kentaro Fukuchi et al 2002) pays its attention to the information that is shown on the Internet in order to extract the background information of companies, however, these research could not collect the community information because they form community while paying attention to the link structure on the Web. Then, this research extracts related information among the companies by analyzing the Web automatically. This is greatly useful for constructing JV. And, we help to construct JV by showing the reliable information among the companies.

AN ABSTRACT OF THIS RESEARCH

This suggested method acquires the related information among the companies on the Web. A figure 1 shows a flow of four processing.

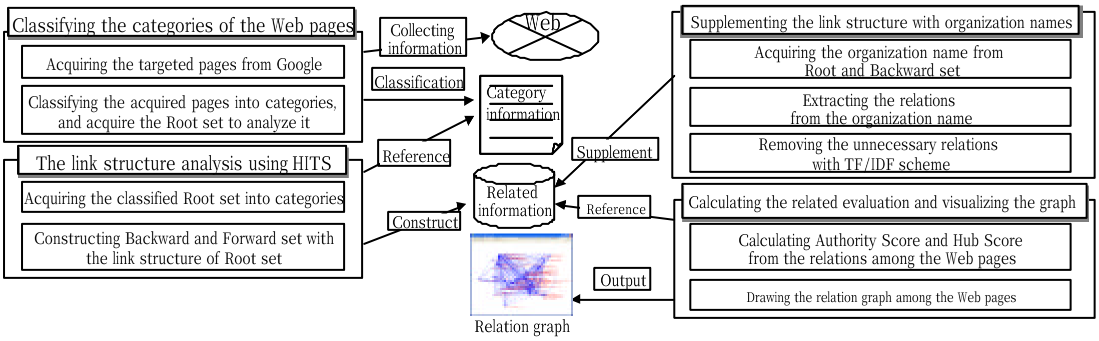
Classifying the categories of the Web pages
Acquiring the targeted pages from Google
Classifying the acquired pages into categories, and acquire the Root set to analyze it
The link structure analysis using HITS
Acquiring the classified Root set into categories
Constructing Backward and Forward set with the link structure of Root set
Collecting information
Web
Classification
Category information
Reference
Construct
Related information
Supplement
Reference
Output
Relation graph
Supplementing the link structure with organization names
Acquiring the organization name from Root and Backward set
Extracting the relations from the organization name
Removing the unnecessary relations with TF/IDF scheme
Calculating the related evaluation and visualizing the graph
Calculating Authority Score and Hub Score from the relations among the Web pages
Drawing the relation graph among the Web pages

Figure 1: Process of the flow

At first, we acquire the Web pages of companies from the Internet in order to categorize them. Next, we classify the acquired pages into categories with the vector space method (Masanori Ikebe et al 2005). There are two types of classification systems. The first is for the treatises such as Universal Decimal Classification and Nippon Decimal Classification. The second is for the Web pages such as category information that is provided by the each portal site on the Internet. However, these systems are not specialized in the construction industry community because they are classified as a general-purpose mark. Then, this research adopts the item that agrees with a rate provided by the cities, towns and villages in order to classify the Web pages into forms useful for JV. And, the Web pages classified into each category are defined as the Root set. This set is used for the standard of the link structure analysis by HITS (Hyperlink-Induced Topic Search) algorithm. The link structure analysis by HITS starts analyzing from the Root set and adds linked Web pages into standard set as Backward set and link Web pages as Forward set. In order to supplement the link structure analysis with the organization name, we extract the organization name from the Web pages using the morphological analysis and the organization name dictionary. And, we calculate the importance degree of the extracted organization name. We judge the related strength in order to delete the weak relations and add the useful relations to the Root set. When we make the graph of the related evaluation, we calculate the evaluation using the Web pages and make the graph based on it.

As a constitution of this thesis, we comment on the category classification of Web pages in Chapter 3. In Chapter 4, we point it out about the traditional link structure analysis and the problems. In Chapter 5, we make an experiment to see the accuracy of the suggested method. Then, in Chapter 6, we describe results and consideration about the prospects.

CLASSIFYING THE WEB PAGES INTO CATEGORIES

In order to classify the Web pages into categories, we compare characteristic documents prepared each categories with new additional Web pages and calculate similar degree. We extract the former documents from the Web pages classified by the portal sites beforehand. In case of calculating similar degree, we use TF/IDF (Term Frequency-Inverse Document Frequency) scheme for each documents and create the feature vector. In the same way, we create it for each category. Then, we open the created feature vector into the multidimensional space and calculate the cosine correlation value between documents and categories. And, we calculate the cosine correlation value for each category and define the greatest value category as the classified category. In an existing category classification research (Hideto Kazawa 2004), a multiple topics problem, one Web page belongs to some category, exists, however, in this research, it is possible that the relation becomes the same in the each category if the companies are classified some category. Then, the purpose of this research is to classify the Web pages into one category.

CREATION OF THE FEATURE VECTOR

We create the feature vector based on the extracted vocabulary that is from the characteristic Web pages and categories. In order to create the feature vector, at first, we carry out the morphological analysis for the targeted documents and extract the noun group in them. Then, we define the noun group extracted from the document *x* as $Kx=\{Nx_1, Nx_2, Nx_3,..., Nx_n\}$. And we define the noun group extracted from the category *y* as $Hy=\{Ny_1, Ny_2, Ny_3,..., Ny_m\}$. In this suggested method, we use ChaSen as the morphological analyzer. ChaSen define a part of speech that is taken the logarithm of appearance probability as the risk and adopts the smallest risk combination among the word line, the part of speech lines that can be realized as the result. Next, we calculate the importance in the document for an extracted noun. In this research, we adopt TF/IDF scheme used as the importance calculation generally. The importance calculating expression by TF/IDF scheme is as follows.

$$TF(Nx_n) = \frac{C(Nx_n)}{\sum_{n'} C(Nx_{n'})} . \qquad (1)$$

$$IDF(Nx_n) = \log \frac{\sum_{x'} x'}{df(C(Nx_n))} . \qquad (2)$$

$$W(Nx_n) = TF(Nx_n) \times IDF(Nx_n) . \qquad (3)$$

The first, in an expression (1), we calculate the weight coefficient of any noun in the Web page. Concretely, we divide any noun NKx_n in document *x* by the total number of noun in the Web page. In an expression (2), we calculate the logarithm next and, adjust the weight of targeted noun using the total number of *KH(x)* and any noun. The *df* shows the total number of the Web pages including any noun here. By this processing, the each weight of the general noun decreases. The last, in an expression (3), we calculate the importance of targeted noun in the Web page by the product of the weight coefficient. $W(Nx_n)$ shows the TFIDF value of the noun Nx_n here. And, we define the feature vector from the importance of each noun in the document. The feature vector created by document *x* is $Vk_x=\{W(Nx_1), W(Nx_2), W(Nx_3),...,W(Nx_n)\}$ here. And the feature vector extracted from category *y* is $Vh_y=\{ W(Ny_1),W(Ny_2),W(Ny_3),..., W(Ny_m)\}$.

THE JUDGMENT OF SIMILAR WEB PAGE AND CATEGORY DEGREE

In order to judge the similar degree of the Web pages and each category, we open the combination of the Web page and the category into the *D(x,y)* dimensional virtual space. And, we define *D(x,y)* as the number of set acquired by an expression (4) comparing the Web page *Kx* with the category *Hy*.

$$D(x, y) = (Kx \cup Hy) \quad . \qquad (4)$$

Then, the feature vector opened the virtual space need to own the *D(x,y)* elements, however, Vk_x, Vh_y, and the defined feature vector can only have *n* number of values, where $D(x,y) \geqq n$. Then, in this research, we supplement the characteristic elements as 0 if the feature vector value of the document does not exist in order to make the dimension number the same. Next, we calculate the cosine correlation from the feature vector in order to calculate the similar degree between documents and categories. The cosine correlation value is calculated from an expression (5).

$$same(Vk_x, Vh_y) = \frac{W(Nx_1)\times W(Ny_1) + W(Nx_2)\times W(Ny_2) + \ldots + W(Nx_n)\times(Ny_n)}{\sqrt{W(Nx_1)^2 + W(Nx_2)^2 + \ldots + W(Nx_n)^2} \times \sqrt{W(Ny_1)^2 + W(Ny_2)^2 + \ldots + W(Ny_n)^2}} \quad . \qquad (5)$$

In the expression (5), the numerator of the right side is the inner product of the feature vector of the document *x* and category *y*. And, the denominator of the left side is the product of the distance between the feature vector and the origin. The cosine correlation value is calculated as an angle between the feature vectors of the documents. And, the result is calculated between 0 and 1. The classified two documents are similar if this result is near to 1. When the calculation of the cosine correlation value completed in the combination of every category, we consider the largest cosine correlation value to be the classified category. Then, we use the Web page classified into each category for the Root set $R(x_1)$ that is based on the analysis.

ACQUIRING THE ORGANIZATION RELATIONS BY THE NATURAL LANGUAGE PROCESSING

SUPPLEMENTING THE RELATED INFORMATION OF THE WEB PAGES

In this research, we do not acquire Forward set in order to restrain Topic Drift as improved HITS algorithm, however, there are the problems that we do not make the satisfactory relation map because of the related information shortage. Therefore, in order to supplementing the information related to the Web pages, we extract the organization names from the sum of set *S(x)* consisted of the Root set and Backward set, and we extract the new related information from matching each titles of Web pages and them. The method of extracting the related information is as follows.

1. Extracting the title, the word of links and documents from the HTML source removed tags beforehand in the set *S(x)*.
2. Using the morphological analysis for the extracted documents, and comparing the organization name dictionary with them in order to extract the same noun as the organization name. Then, we define the extracted organization name as $T_1(xy_n)$ and its number as $C(xy_n)$, where *y* is the extracted organization name.
3. Acquiring the new set of the Web pages as the candidate set in order to relate the organization name with them. And defining the sum of this and *S(x)* as $S_{add}(x)$.
4. Dealing with the morphological analysis for the each title of $S_{add}(x)$. And, extracting the agreed noun as the organization name from comparing the extracted noun with the organization name dictionary. Then, defining extracted the organization names as $T_2(xy_m)$.

5. In order to define agreed relations as $RI_{nl}(xy_i)$, matching $T_1(xy_n)$ with $T_2(xy_m)$.
6. In $RI_{nl}(xy_i)$, in order to remove the weak relations, calculating the each weight coefficient of the organization name using TF/IDF scheme for the importance of the words in the documents.
7. In the each weight coefficient of the organization name, extracting the available information by cutting of the threshold.

In order to extract the organization name in this suggested method, we also use the ChaSen as the morphological analyzer. The reason why we deal with morphological analysis before extracting the organization name is to avoid the word except the noun that is misjudged as the organization in case of matching the Web page contents with the organization name dictionary simplistically. And, we used the organization name dictionary included ChaSen in order to extract the organization name. The number of words that is recorded by it is 16610 cases. And, we also adopt TF/IDF scheme in order to calculate the importance of the each organization name. At first, in the expression (1), we divide any organization name y_n in the Web page x by the total number of the organization name in order to acquire any weight coefficient in the Web page. Second, in the expression (2), we calculate the logarithm and adjust the weight of the organization name using the total number of $S_{add}(x)$ and any words, where df shows the total number of the Web pages including any organization name. With this treatment, the weight of the famous organization per one organization name decreases. Last, in the expression (3), we calculate the importance of the extracted organization name in its Web page using the product of the weight coefficient. Then, we calculate the threshold in order to cut off the related information among the each organization. We calculate the threshold, at first, we make the correct answer from extracting the relations among the organizations of any number of Web pages in set $S_{add}(x)$ by the hand. Next, we compare the correct answer with the related information acquired by the suggested method and calculate an F measure. We define it as evaluation value and calculate the most suitable threshold using the steepest descent method. The F measure is the method that uses in the evaluation of extracting the specific words. It is defined as the multiplied product of the recall by precision. Now, the former is the correct answer rate of extracted data and the latter is the supplement rate of the correct answer. When calculating the threshold, we extract the random initial value and try several times in order to avoid the local minima problem. Therefore, we remove the weak relations and improve the quality of the related information.

CALCULATING THE RELATED EVALUATION VALUE

Authority Score and Hub Score in HITS algorithm is greatly reliable evaluation system because a lot of search engine and existing research use it today. Then, in this research, we also calculate the related evaluation value by Authority Score and Hub Score as well as HITS algorithm. The method of calculating the evaluation value is as follows.

1. We extract the title, the word of links and documents from the HTML source removed tags beforehand in the set $S(x)$.
2. Acquiring the total set of the related information, $RI_{all}(xy_i)$ by the sum of $RI_{back}(x)$.

3. Calculating Authority Score and Hub Score from $S_{add}(x)$ and $RI_{all}(xy_i)$ using HITS algorithm.

VISUALIZING THE RELATED INFORMATION

After calculating Authority Score and Hub Score, in this suggested method, we remove the relations from $S_{add}(x)$ that their Authority Score converged in 0. And we define the left set of the Web pages as $S_{auth}(x)$. And, in calculating the $S_{auth}(x)$, we also remove the Web pages and related information from $RI_{all}(xy_i)$. Then, we draw the graph $G(x)$, where $S_{auth}(x)$ is the apex and $RI_{all}(xy_i)$ is the side of it. When we draw the graph $G(x)$, we put $S_{auth}(x)$ on the latticed plan with the equal interval in random. And in the each apex, we fix the placement by calculating the balance of the entire apex that is connected by the related information. After that, we draw the side between the apexes using the related information.

EVALUATION EXPERIMENT

PRECISION EVALUATION OF THE CLASSIFYING THE CATEGORIES OF THE WEB PAGES

In this session, we classify Web pages into categories in order to evaluate their precision. We adopt the classification category that is the same between the local public entity and the portal site on the Internet, Yahoo! Japan. Concretely, we choose the 8 categories such as the civil engineering work, building, pipe work, painting, paving work, landscape gardening, electrical work and water supply and sewerage. And we adopt the Web site document information of the companies as a training sample that is categorized in Yahoo! Japan. The Web pages set that will be classified in this experiment are adopted from a result searched by the key word that combined “co.jp” with each category name. We use Google for a search engine. As a result, the civil engineering works categories are 360 cases, the building are 112, the pipe work are 88, the painting are 96, the paving work are 176, the landscape gardening are 120, the electrical work are 112 and the water supply and sewerage are 152. In order to evaluate the category classification, we check the result by hands and calculate the each categories of the F measure. The F measure of the each category is as follows.

Table 1: Evaluation of the classified categories using the F measure

	civil engineering work	building	pipe work	painting	paving work	landscape gardening	electrical work	water supply and sewerage
Precision	0.95	0.68	0.47	0.42	0.79	0.68	0.47	0.79
Recall	0.42	0.93	0.82	0.67	0.68	0.87	0.64	0.79
F-measure	0.58	0.79	0.60	0.52	0.73	0.76	0.55	0.79

By the result of the table 1, the classified category led by the suggested method is seems to be leant because the average of the F measure is 0.66. This is because the training sample from the Yahoo! Japan does not include the vocabulary that shows the characteristic key word. And in case of the civil engineering works category, there is a tendency to overlap some of the categories. It is because the civil engineering work also includes the characteristic of the paving work and the pipe work. Therefore, the civil engineering work extracts three times relations compared with other categories and is the low precision.

PRECISION EVALUATION OF THE CLASSIFYING THE CATEGORIES OF THE WEB PAGES

In this session, in order to evaluate the related information, we extract the available Web pages from the Web pages of the each organization categories. And we evaluate the quality of each relation. In this experiment, we acquire the quality of the precision evaluated by the F measure.

We define the acquired Web page that is classified into categories as the Root set. And, in this experiment, we pay the attention to the civil category because there are a lot of companies that have a qualification. And in this suggested method, we define the 1452 cases that combined the 360 cases Root set with the 1092 cases Backward set as the candidate set using the link structure analysis. Then, we carry out the link structure analysis. We are able to extract the 835 relations from the collected Web pages. In order to evaluate the quality of the related information, we compare these acquired related information with the correct related information that has 1029 cases extracted from the Root set. Then, we acquire the 0.38 F measure. In order to extract the available relations, we define the 0.0314332 as the threshold. This value is calculated by the random acquisition of the initial value and using the steepest descent method.

At last, we are able to acquire the 0.59 F measure. The reason why the F measure stays 0.59, there are less companies that have their own Web page in the construction industry community than other industry.

VISUALIZING THE RELATED ORGANIZATION GRAPH

In this session, in order to confirm the Web community from the acquired relations, we visualize (Konomu Dobashi 2003) the related organization graph. The way to visualize it, at first, we define the one Web page of the organization as the basic point and extract the Web communities from its around. And we remove the Web pages that converged in 0 an extract the main communities. We put the Web pages on the random position and draw these relations as the line. And we set the each point on the place calculated the balance and visualize the result of the suggested method by a figure 2.

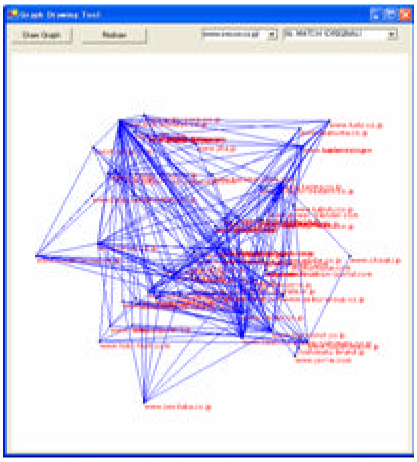

Figure 2: Visualizing the related information using the natural language processing

In the figure 2, the graph shows the related organization about www.smcon.co.jp (*Sumitomo Mitsui Construction Co., Ltd*). The result shows that www.smcon.co.jp has direct relations with www.shimz.co.jp (*Shimizu Corporation*) and const.tokyu.com (*Tokyu Construction*). And through the one organization, it connects www.obayashi.co.jp (*Obayashi corporation*). These companies actually work together as the specificity construction joint venture of the *Shimizu*, *Obayashi*, *Tokyu* and *Mitsuisumitomo*, and we are able to extract the correct relations among the companies. Moreover, this suggested method extracts the new candidate in JV construction such as www.penta-ocean.co.jp (*Penta-Ocean Construction Co., Ltd*) and www.hitachi-cement.co.jp (*Hitchi-Cement Co., Ltd*). Therefore, in the method, we are able to extract the available relations in the target category and offer one judgment materials in JV construction.

CONCLUSION

On this research, we succeeded in extracting background information about the target company that became important when constructing JV. The background information that was able to be acquired by this suggestion proved to be the relations with having built JV at the past, tie-up relations and dealings. This information is able to distinguish the reliability of the target companies and the result of the joint enterprises by combining the business results and evaluation information. For example, it is possible to confirm the degree of contribution by combining constructed JV at the past and its evaluation information. We performed experiments in the creation of the related information specialized construction industry in order to help to construct JV. Using the same way, we expect to extract the related information among the technical term dictionaries from the Web by changing the organization names with the technical terms. However, this suggested method has the purpose of classing organization names into the only one category. Then, the method cannot expect enough classification precision when a company that diversifies of business contents uses the system because it is difficult to identify the type of industries. And, the method cannot expect enough precision when targeting on a general word because we use important words in the Web page for creating the related information. Then, we help to construct JV among the different type of industries by dealing with multi-topic text categorization. And, we are going to make full use of RDF (Resource Description Framework) on Semantic Web and FOAF (Friend of a Friend) in order to extracting more reliable related information among the companies.

REFERENCES

Alice Nakamura and Masao Nakamura (2004). "Firm Performance, Knowledge Transfer and International Joint Ventures." *International Journal of Technology Management, Inderscience Publishers*, 27 (8) 731-746

Chen Chung-Jen (2004). "The Effects of Knowledge Attribute, Alliance Characteristics, and Absorptive Capacity on Knowledge Transfer Performance." *R&D Management*, 34 (3) 311-321

Fusashi Nakamura and Hideyuki Mizuta (2004). "Computational Organization Study on Enterprise Hierarchy and Communication Network." *Information Processing, Information Processing Society of Japan*, 45 (9) 950-955 (in Japanese)

Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda (2004). "Maximum Margin Labeling for Multi-Topic Text Categorization." *Special Interest Group of Natural Language Processing Notes of IPSJ*, *Information Processing Society of Japan*, 2004 (NL-163) 53-60 (in Japanese)

James Bamford, David Ernst, David G. Fubini (2004). "Launching a World-Class Joint Venture." *Harvard Business Review*, *Harvard Business School Publishing*, 82 (2) 90-100

Kentaro Fukuchi, Masashi Toyoda, and Masaru Kiretsugawa (2002). "Implementation of Web Community Browser and its evaluation." *The Institute of Electronics and Information and Communication Engineers Technical Report*, *The Institute of Electronics and Information and Communication Engineers*, 102 (209) 79-84 (in Japanese)

Konomu Dobashi, Hiroyuki Yamauchi, and Ryuki Tachibana (2003). "Key Term Extraction and Visualization for Knowledge Chain Discovery Support-Visual function of TermLinker system-." *Special Interest Group on Intelligence and Complex Systems Notes of IPSJ*, *Information Processing Society of Japan*, 103 (304) 41-46 (in Japanese)

Masanori Ikebe, Shigenori Tanaka, Hitoshi Furuta, and Kenji Nakamura (2005). "Research of Document Data Identity Determination Component for Electronic Delivery." *Journal of Civil Engineering Information Application Technology*, 14 7-14 (in Japanese)

Ministry of Land, Infrastructure and Transport (2005). Construction and the reason why duty added to organize joint venture in the specific construction (in Japanese)

Miyazaki Prefecture (2005). Connection manual among the companies for construction industry (in Japanese)

Jiro Shibano (2004). "Merits of Implementing the Balanced Scorecard and the Points to Consider--Based on the Experiences of the BSC Implementation Study Team--." *Special Interest Group on Information Systems Notes of IPSJ, Information Processing Society of Japan*, 2004 (53) 9-12 (in Japanese)